

# OLAP Server: Focusing on Performance

**Jeff Polzin, Barbara Walters, Vicki Jones, and Bob Janka**

*Abstract: Beginning with SAS 9, the SAS Online Analytical Processing (OLAP) Server began to concurrently process multiple user requests to support multi- user access to OLAP Data. This new server technology has added to the complex task of server performance tuning and system performance analysis. The intricate interaction of hardware, network, and software components warrants a methodical approach to monitoring, debugging, and improving performance. In this paper we provide a technical overview of the architecture of the SAS OLAP Server and a practical approach to identifying performance issues and effecting improvement. Discussion will include both general guidelines and specific experiential recommendations.*

## **Introduction**

The SAS Intelligence Value Chain (IVC) brings component capabilities together into a single, unified framework that defines the requirements for enterprise intelligence adoption. As each element of the Intelligence Value Chain is tightly integrated, organizations realize successive, derived value. Operational data readily becomes enriched, cleansed and trustworthy information that is surfaced to deliver new insights, provide answers and drive strategy. The result? Intelligence that helps companies deliver current business results they need and confidently predict future business requirements. It's this synergistic approach that makes the difference between general, isolated pockets of information efficiencies and consistent, usable, enterprise-wide effectiveness – this is the SAS advantage.

The components of the SAS Intelligence Value Chain are:

- Plan - A set of proven, best-practice, roadmaps that are supported by integrated industry data models, project methodologies and consulting expertise that reliably create and deploy customized, high-benefit Solutions.
- ETL<sup>Q</sup> raises the typical function of “ETL” with the power of Data Quality. SAS ETL<sup>Q</sup> is an integrated ETL platform that synthesizes corporate data from operational silos of information on any platform and in any format.
- Intelligent Storage is a dedicated platform designed from the outset to efficiently disseminate information for both business intelligence and analytic requirements.
- Business Intelligence is a set of out-of-the-box, enterprise-wide ad-hoc query and reporting capabilities that focuses on usability, mass distribution, and consumption of information – all of which are designed to maximize the access, exploration and presentation of data for informational insight.

- Analytic Intelligence is an integrated platform for exploring and leveraging enterprise data to create descriptive and predictive information that provides insight, guidance, and support for better decision making.

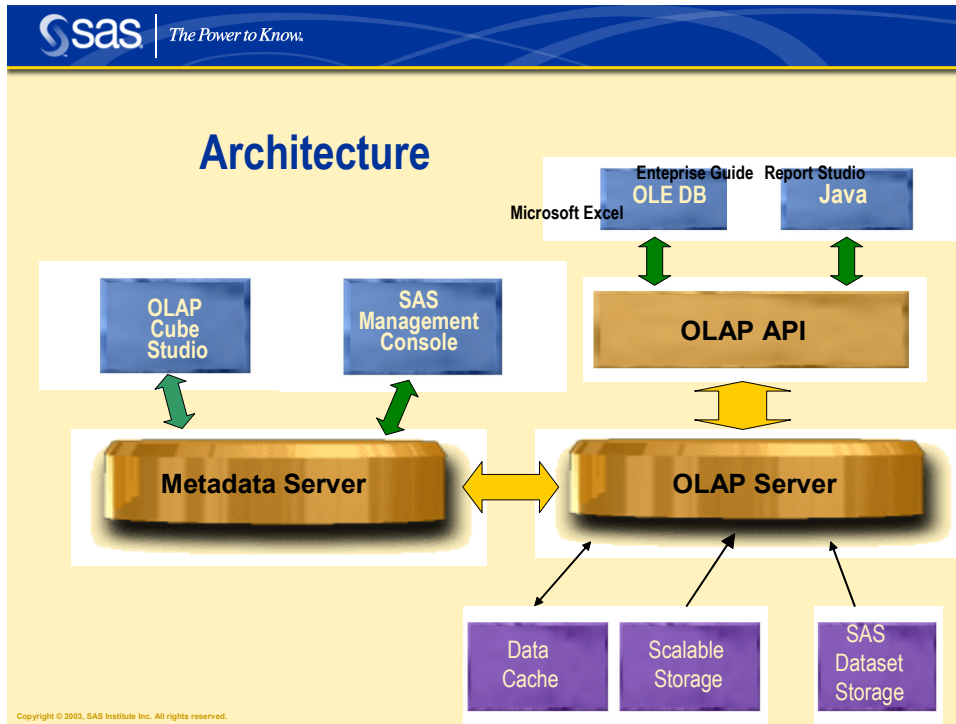
Through fast, efficient access to summarized data, SAS Online Analytical Processing (OLAP) Technology enables the multidimensional analysis capabilities at the core of SAS Enabling Technologies and gives customers complete control over global views of their business.

Decision makers, regardless of computing expertise, can view business scenarios from a number of perspectives. You can produce anything from standard tables, charts and maps to advanced multidimensional reports that might include data visualization and geographical analysis. Drill down and across data views. Take advantage of hot-spotting and traffic-lighting capabilities to identify business trends and long-term developments. Pinpoint critical success factors and key performance indicators to focus on key management issues.

In this paper we focus on performance of this new SAS OLAP technology. We discuss two primary metrics of performance: query response time and number of transactions processed per second (i.e., throughput). After a brief introduction, we outline the critical factors for performance, discuss approaches for monitoring and debugging performance issues, and system tuning options. We conclude with some general advice on getting the most out of your SAS OLAP server.

## ***OLAP Server Overview***

In SAS 9, the SAS OLAP Server has been rewritten to use the SAS scalable architecture to support high performance scalable multidimensional analysis. Our new multi-user, scalable server component natively supports industry standards such as OLE DB for OLAP and the MDX query language. The SAS OLAP Server software architecture is depicted below.



Users interact with the SAS OLAP server in two ways. First, to create and manage OLAP cubes using the OLAP Cube Studio and SAS Management Console. These application front-ends to the SAS Open Metadata Server serve to define the input data relationships and the definition of a cube's organization and security. Secondly, the SAS OLAP server provides services to the SAS Business Intelligence and Analytical Intelligence applications to provide information to meet the requirements of a specific end user need. Industry standard interfaces are available to service end user needs when using third party query and reporting tools within an enterprise organization.

## Scalability

In information technology, scalability can be defined in two manners depending on the view of the application.

1. It is the ability of a computer application or product (hardware or software) to continue to function well when it (or its context) is changed in size or volume in order to meet a user need. Typically, the rescaling is to a larger size or volume. The rescaling can be of the product itself (for example, a line of computer systems of different sizes in terms of storage, RAM, and so forth) or in the scalable object's movement to a new context (for example, a new operating system).
2. It is the ability not only to function well in the rescaled situation, but to actually take full advantage of it. For example, an application program would be scalable if it could be moved from a smaller to a larger operating system and take full

advantage of the larger operating system in terms of performance (user response time and so forth) and the larger number of users that could be handled.

It is usually easier to have scalability upward rather than downward since developers often must make full use of a system's resources (for example, the amount of disk storage available) when an application is initially coded. Scaling a product downward may mean trying to achieve the same results in a more constrained environment. The SAS OLAP server will provide support for both of these situations as required by the end user.

## ***Critical Factors for Performance***

When looking for performance improvements, most environmental components should be considered. Here we identify three factors---hardware platform, cube characteristics, and query characteristics---and discuss their impact on performance.

### **Hardware Platform**

As expected, query response time and system throughput are heavily dependent on the host hardware and operating system. With SAS 9, new technology has been implemented to exploit available hardware and operating system capabilities. In particular, multi-threaded technology is exploited to take advantage of additional processing power available on multi-processor systems.

The primary system factors that impact performance are the number and speed of processors; bandwidth and speed of I/O subsystem; and amount and speed of the memory subsystem. Given multiple processors, new multithreaded technology enables simultaneous processing of multiple concurrent simple queries and complex CPU-bound queries. Tasks required to complete each query are distributed between multiple threads, allowing a single-user (i.e., non-concurrent) workload to take advantage of multi-processor based systems.

For workloads with many concurrent, simple queries for which the data all fits in memory, we expect query response time to remain constant and throughput to increase proportionately with the number of processors, up to the number of processors. Once all processors become saturated, queries queue, response time increases to include query wait time, and throughput remains relatively constant.

The speed of processors is most **important for** CPU bound workloads. On systems where the “working set” of cube data does not fit in memory, the speed of the I/O subsystem becomes a factor. For example, processing a workload which constantly accesses multiple cubes and their associated aggregations that do not all fit in memory will result in each query waiting until its data is retrieved from disk, thus impacting both response time and throughput. When selecting a hardware platform to host a SAS OLAP server the selection of the disk I/O controllers and drives is important. With the continuous advances in high performance disk I/O technology, such as SCSI and Serial

ATA, is important to choose the technology that will meet your expected needs and budgetary requirements.

When memory contention is high, operating system paging activity can reduce the performance of the OLAP server. If the disk space used for operating system paging is on the same disk configuration or controller as the cube data, this configuration can create bottlenecks in the I/O subsystem due to the high contention for the same I/O pathways. Moving the paging disk space to a separate disk drive or controller will reduce the I/O pathway contention and reduce the impact on the OLAP server and as it accesses cube data on behalf of its connected clients.

As implied above, the mix of queries in a typical workload affects performance. Typically, the best performance will be achieved when all queries are against the same *cached* data. As the data requirement's for each query varies, so does the memory footprint of the workload, and the amount of memory required to "cache" the workload.

### **Cube Characteristics**

The size of the cube (N-way) depends on the number of dimensions, the number of hierarchies in each dimension, the number of levels in each hierarchy, the cardinality of members at each level, and the number of measures to be computed. The larger the N-way, the more memory necessary to keep it all cached. The primary influence on the size of the cube is the cardinality of members at each level and the number of measures.

In addition to the N-way itself, supplemental files such as metadata, indexes, and aggregation files are built. While the metadata and index files are relatively small, the N-way and the additional aggregations can be quite large, making the tradeoff between query time speed, cube build time processing, and continual disk usage an interesting dilemma. When query response is critical the consideration for creating additional aggregations should be considered. With the additional aggregations in place the SAS OLAP server has to perform fewer rollup operations to summarize lower levels of a cube's dimension.

Cube partitioning options are available when an administrator creates the definition of an OLAP cube. These options allow the designer to distribute the both the cube's aggregation data and associated indexes across multiple disk storage devices or controllers. In large cube installations, the effects of partitioning data and indexes across multiple disk controllers should be considered to optimize the host platforms I/O subsystem to perform parallel I/O without serialization in the controller or on the devices connected to a physical disk controller.

### **Query Characteristics**

Finally, the workload itself is an important factor in performance. When all data is cached, simple queries to retrieve data points either from the N-way itself or an aggregation will have the best response time and peak throughput. Queries that require additional computation will necessarily take longer. That is drilling (up or down) and

aggregating on the fly will require potentially much more data to be loaded (if not cached) and more computation during query execution.

Reviewing the types of MDX queries being processed can assist in tuning the performance of the OLAP server by indicating the need for additional aggregations and platform resources or modifications to current cube definitions and platform resources.

## ***Monitoring and Debugging***

To determine the performance characteristics of an SAS OLAP installation running under a customer site's typical user load requires the capability to analyze system specific performance measurements on a properly configured host platform. Typically performance metrics are maintained by the operating system for system administrators or performance analysts to monitor and make appropriate adjustments to the operating system, tune running applications, or to adjust the configuration of the hardware platform itself.

Specifically, the typical measurements to study relate to the utilization of the processor's resources, memory management and paging characteristics, I/O subsystem performance characteristics, network I/O activity between all component's, and the operating system's internal processing.

In a heavily loaded SAS OLAP Server environment, the CPU utilization should be very high across all processors installed on the host platform. Sustained low CPU utilization is typically an indication of a high level of resource consumption by other components in the environment such as higher memory or I/O utilization. Under high memory contention situations there may be other applications that are co-located on the host platform that may be interacting with each other in detrimental ways by increasing the paging requirements due to reduction in available memory between the applications.

Measurements that reflect I/O performance should be closely monitored to insure that client's and cube data is being transferred efficiently between the disk storage devices and main memory. Increases in elapse time for end-user queries tends to indicative of I/O wait state situations.

Another tuning methodology that can be used to monitor the performance of a running SAS OLAP server is to enable the built-in logging capabilities. Using the industry standard Application Response Measurement (ARM) format, log records are generated indicating query content and query start and completion times. From this data, information regarding aggregation usage, individual query response times, or throughput can be determined. If query response times increase or are initially too high, the ARM data may indicate that a cube's N-way table is being accessed too often, as required in performing lower level rollups of a dimension. In these situations, the administrator should review the cube's currently defined aggregations to determine if building

additional aggregations or modifications of existing aggregations at cube creation time would be beneficial.

## ***Tuning Options***

Within the SAS OLAP server there are several tuning options that should be reviewed to insure optimal settings for the server configuration.

The SAS ***MEMSIZE*** option informs the SAS system running the OLAP server to limit that virtual address space of the server. This is a maximum threshold honored by the server. Once this memory size is exhausted user connections and MDX queries will begin to fail due to lack of available resources required for processing.

The SAS ***REALMEMSIZE*** option informs the server how much physical memory it is allowed to consume. It is recommended that this option be set to the actual physical memory size that is installed on the host platform. The standard default value is 96MB which is too low for most server operations.

Other OLAP Server based tuning options are available during the setup of the server via the SAS Management Console. Also a number of OLAP server specific and cube specific tuning options are available via the Cube Design Studio to instruct the server: to enable caching of aggregation results, establish cache buffer characteristics, and to define thread utilization thresholds, among others.

## ***Conclusions***

With the SAS Intelligence Value Chain (IVC) you get a single, unified framework to support enterprise intelligence. Integral functionality of the IVC includes dissemination of information for both business intelligence and analytics. SAS OLAP technology and been designed and optimized to meet this need. The SAS OLAP server provides fast, efficient access to summarized data, enabling multidimensional analysis and control over customizing views of information.

Important factors to consider in achieving the best performance from your OLAP server include the hardware, cube design and implementation, and query formulation. During the planning stage careful consideration of the expected workload can aid in identifying potential scarce resources (e.g., processing power, I/O or network bandwidth) and allow pre-emptive resolution of performance bottlenecks. Knowledge of the workload will also be invaluable in making decisions regarding cube design and location of cube data on disk. Once implemented, several tuning options are available to aid in improving performance.

Along with careful design of your OLAP cube and queries, continued monitoring of system resource utilization will assist you in getting the most out of your OLAP server.

At SAS, research and development teams continue to be committed to provide exceptional functionality with optimized performance.